EXHIBIT X

The Empirical Distribution Function with Arbitrarily Grouped, Censored and Truncated Data

By BRUCE W. TURNBULL

University of Oxford and Cornell University

[Received May 1975. Revised May 1976]

SUMMARY

This paper is concerned with the non-parametric estimation of a distribution function F, when the data are incomplete due to grouping, censoring and/or truncation. Using the idea of self-consistency, a simple algorithm is constructed and shown to converge monotonically to yield a maximum likelihood estimate of F. An application to hypothesis testing is indicated.

Keywords: EMPIRICAL DISTRIBUTION FUNCTION; SURVIVAL CURVE; CENSORING; TRUNCATION; GROUPING; MAXIMUM LIKELIHOOD; KAPLAN–MEIER PRODUCT LIMIT ESTIMATOR; SELF-CONSISTENCY; NEWTON–RAPHSON; MULTINOMIAL DISTRIBUTION; LOGRANK TEST

1. INTRODUCTION

WE consider the non-parametric estimation of the distribution function F of a real-valued random variable X, when the sample data are incomplete due to restricted observation brought about by grouping, censoring and/or truncation. More precisely the situation is as follows. Subsets $B_1, B_2, \ldots, B_N$ of the real line are given and there are N independent observations $X_1 = x_1, \ldots, X_N = x_N$, where $X_i\,(1 \leqslant i \leqslant N)$ is drawn from the truncated distribution $F(x;\, B_i) = P(X \leqslant x \mid X \in B_i)$, $X \in B_i$. Thus X_i is *truncated* by B_i or, in other words, the experimenter would not have been aware of the existence of that observation had X_i not belonged to B_i. Moreover $X_i\,(1 \leqslant i \leqslant N)$ may not be observed exactly and is known only to lie in the set A_i where $A_i \subseteq B_i$. Thus X_i is censored into the set A_i. Grouped data can be naturally considered as censored, where each observation is censored into one of a fixed collection of disjoint sets. The observed data are then the N pairs $(A_1, B_1), (A_2, B_2), \ldots, (A_N, B_N)$.

The truncating sets $\{B_i\}$ can either be viewed as fixed or as random. We can now think of a partition of the set B_i and A_i as that member of the partition into which X_i falls. Again the partition can be viewed either as fixed or as having arisen from some random mechanism independent of X_i. In many cases, the partition of B_i will be unknown (except for the fact that A_i belongs to it); these assumptions will make knowledge of the partition irrelevant to the estimation of F. The case of grouped data can be considered as one in which the partitions are known and are the same for each $i\,(1 \leqslant i \leqslant N)$.

If $B_i = (-\infty, \infty)$ then X_i is not truncated, and if A_i consists of a single point then X_i is uncensored, i.e. is *exact*. We say that X_i is *interval* censored if A_i is of the form $[L_i, R_i]$ and X_i is *right* (*left*) censored if $R_i = +\infty$ ($L_i = -\infty$). Of course if $L_i = R_i$, then X_i is exact. Interval, right and left truncation are defined similarly.

Interval censoring occurs naturally when the $\{X_i\}$ represent response times and there is periodic inspection, e.g. in medical or correctional follow-up or in industrial life-testing. Here the sets $\{A_i\}$ may overlap because sample items can enter the programme at different ages. Also the bioassay problem of Ayer *et al.* (1955) can be considered an extreme case of double censoring. Truncation can occur if the population from which X_i is drawn has been subject to some screening procedure in which all items with x-values outside B_i have been removed. This situation can arise in consumer product testing, for example. Concerning survivorship analysis, Mantel (1966) mentions left truncation in the context of merging clinical trials. Here a group of survivors at a certain point in time is to be incorporated into ongoing

study data when the original size of the group of which these are a remnant is unknown. The re-entry problem is another example where there can be a more general truncation pattern. This situation occurs when a person can be lost to follow-up, by leaving a health insurance programme for instance, but then he rejoins at a later date.

If some parametric form for F can be assumed, then the method of Blight (1970) can be applied directly. (See also Hartley and Hocking, 1971 and Sundberg, 1974.) We will assume that there are no parametric assumptions and obtain the analogue of the sample c.d.f. (when there is exact data with no truncation) and of the Kaplan and Meier (1958) PL estimator (right censoring with possibly left truncation).

2. Reduction of the Problem

We first show that the maximum likelihood estimate, $\hat{F}$, of F increases in only a finite number of disjoint intervals (or points). We shall use the same notation as Peto (1973) who obtained a similar result for interval censoring with no truncation.

Let us assume that each A_i $(1 \leqslant i \leqslant N)$ can be expressed as the finite union of disjoint closed intervals, with the convention that an isolated point $\{x\}$ is a closed interval $[x, x]$ and that a semi-infinite interval is semi-closed only. Thus we can write

$$A_i = \bigcup_{j=1}^{k_i} [L_{ij}, R_{ij}] \quad (i = 1, 2, \ldots, N),$$

where $-\infty \leqslant L_{i1} \leqslant R_{i1} < L_{i2} \leqslant \ldots < L_{ik_i} \leqslant R_{ik_i} \leqslant \infty$ and $R_{i1} > -\infty$, $L_{ik_i} < \infty$. We now construct a set of disjoint intervals whose left and right end points lie in the sets $\{L_{ij};\ 1 \leqslant j \leqslant k_i,\ 1 \leqslant i \leqslant N\}$ and $\{R_{ij};\ 1 \leqslant j \leqslant k_i,\ 1 \leqslant i \leqslant N\}$ respectively, and which contain no other members of $\{L_{ij}\}$ or $\{R_{ij}\}$ except at their end points. We write these intervals $[q_1, p_1], [q_2, p_2], \ldots, [q_m, p_m]$, where $q_1 \leqslant p_1 < q_2 \leqslant \ldots < q_m \leqslant p_m$. Also define $C = \bigcup_{j=1}^{m} [q_j, p_j]$.

Under the assumption of Section 1, the likelihood is proportional to

$$L^*(F) = \prod_{i=1}^{N} \left[\sum_{j=1}^{k_i} \{F(R_{ij}+) - F(L_{ij}-)\} \right] \Big/ P_F(B_i). \tag{2.1}$$

We will assume that $P_F(\cup B_i) = 1$, which occurs for instance if at least one observation is not truncated. The search for that function F that maximizes (2.1) is facilitated by the following two lemmas, which can be easily proved upon examination of L^*.

Lemma 1. Any c.d.f. which increases outside the set C cannot be a maximum likelihood estimate of F, except in the trivial case when $A_i \cap C = B_i \cap C$ for all i.

Lemma 2. For fixed values of $F(p_j+)$, $F(q_j-)$ $(1 \leqslant j \leqslant m)$, the likelihood is independent of the behaviour of F within each interval $[q_j, p_j]$.

Now, for $1 \leqslant j \leqslant m$, define $s_j = F(p_j+) - F(q_j-)$. Then the vectors $\mathbf{s} = (s_1, \ldots, s_m)$ where $\sum s_j = 1$ and $s_j \geqslant 0$, define equivalence classes on the space of distribution functions F which are flat outside C. We will say that two such functions are equivalent if they have the same $\mathbf{s}$-vectors. All functions in the same equivalence class will have the same likelihood by Lemma 2, and Lemma 1 shows that we can restrict our search for an MLE to these classes. Therefore the MLE will, at best, be unique only up to equivalence defined in this way. For example, for right censored data, the Kaplan–Meier PL estimate is undefined at the exact observation points and in an interval $[L, \infty]$, say, if the largest observation is at L and is censored.

The foregoing discussion shows the problem of maximizing (2.1) reduces to one of maximizing

$$L^*(s_1, \ldots, s_m) = \prod_{i=1}^{N} \left(\sum_{j=1}^{m} \alpha_{ij} s_j \Big/ \sum_{j=1}^{m} \beta_{ij} s_j \right), \tag{2.2}$$

subject to $\sum s_j = 1$, $s_j \geqslant 0$ $(1 \leqslant j \leqslant m)$, where $\alpha_{ij} = 1$ if $[q_j, p_j] \subseteq A_i$, 0 otherwise, and $\beta_{ij} = 1$ if $[q_j, p_j] \subseteq B_i$, 0 otherwise. We remark that we would be able to write down (2.2) immediately as the likelihood if there were a discrete scale for X (i.e. X could only take on values

$t_1, t_2, \ldots, t_m$, say). Then we would define $s_j = P(X = t_j)$. Let $\hat{\mathbf{s}} = (\hat{s}_1, \ldots, \hat{s}_m)$ denote a value of $\mathbf{s}$ for which L^* attains its maximum in the region $\mathscr{R} = \{\mathbf{s} | \sum s_j = 1, s_j \geqslant 0\ (1 \leqslant j \leqslant m)\}$. We assume that neither of the following two trivial situations hold: (A) There exist j, k with $1 \leqslant j, k \leqslant m$ and $j \neq k$ such that $\alpha_{ij} = \alpha_{ik}$ for all $i\ (1 \leqslant i \leqslant N)$, nor (B) There exists a subset D such that for each i, $1 \leqslant i \leqslant N$, either $B_i \cap C \subseteq D$ or $B_i \cap C \subseteq D^c$. If (A) occurs, L^* depends on s_j and s_k only through their sum. In case (B) only the ratio $s_j/(\sum_{k \in D} s_k)$ is estimable for $j \in D$ and hence $\hat{s}_j$ is defined only up to a multiplicative constant. If either (A) or (B) occurs, $\hat{\mathbf{s}}$ is not unique and the maximum likelihood estimate $\hat{F}$ will be determined only as far as belonging to a certain union of equivalence classes.

3. The Self-consistency Algorithm

In this section, we describe an algorithm for obtaining the MLE of $\mathbf{s}$ based on the equivalence between the property of maximum likelihood and that of self-consistency. This latter property will be defined precisely below; it is an extension of the idea first used by Efron (1967) for right censored data and later by Turnbull (1974) for doubly censored data.

For $1 \leqslant i \leqslant N$, $1 \leqslant j \leqslant m$, let $I_{ij} = 1$ if $x_i \in [q_j, p_j]$ and 0 otherwise. Because of the censoring the value of I_{ij} may not be known, however its expectation is given by

$$E_{\mathbf{s}}[I_{ij}] = \alpha_{ij} s_j \Big/ \sum_{k=1}^{m} \alpha_{ik} s_k = \mu_{ij}(\mathbf{s}), \quad \text{say.} \tag{3.1}$$

Thus $\mu_{ij}(\mathbf{s})$ represents the probability that the ith observation lies in $[q_j, p_j]$ when F belongs to the equivalence class defined by $\mathbf{s} = (s_1, \ldots, s_m)$. Also, because of the truncation, each observation $X_i = x_i$, can be considered a remnant of a group, the size of which is unknown and all (except the one observed) with x-values in B_i^c. (They can be thought of as X_i's "ghosts".) Let J_{ij} be the number in the group corresponding to the ith observation which have values in $[q_j, p_j]$. Of course J_{ij} is unknown but its expectation, under $\mathbf{s}$, is given by

$$E_{\mathbf{s}}(J_{ij}) = (1 - \beta_{ij}) s_j \Big/ \sum_{k=1}^{m} \beta_{ik} s_k = \nu_{ij}(\mathbf{s}), \quad \text{say.} \tag{3.2}$$

If we treated (3.1) and (3.2) as observed rather than expected frequencies, the proportion of observations in interval $[q_j, p_j]$ is

$$\sum_{i=1}^{N} \{\mu_{ij}(\mathbf{s}) + \nu_{ij}(\mathbf{s})\} / M(\mathbf{s}) = \pi_j(\mathbf{s}), \quad \text{say,} \tag{3.3}$$

where

$$M(\mathbf{s}) = \sum_{i=1}^{N} \sum_{j=1}^{m} \{\mu_{ij}(\mathbf{s}) + \nu_{ij}(\mathbf{s})\}.$$

Note that $M(s) \geqslant N$ with equality if there is no truncation for then $\nu_{ij} = 0$ for all i, j. We say that the vector of probabilities $\mathbf{s}$ is *self-consistent* if

$$s_j = \pi_j(s_1, \ldots, s_m) \quad (1 \leqslant j \leqslant m). \tag{3.4}$$

A *self-consistent estimate* of $\mathbf{s}$ is defined to be any solution of the simultaneous equations (3.4). The form of (3.4) immediately suggests an iterative procedure for finding the solution.

A. Obtain initial estimates $s_j^0\ (1 \leqslant j \leqslant m)$. This can be any set of positive numbers summing to unity, e.g. $s_j = 1/m$ for all j.

B. Evaluate $\mu_{ij}(\mathbf{s}^0)$ and $\nu_{ij}(\mathbf{s}^0)$ for $1 \leqslant i \leqslant N$ and $1 \leqslant j \leqslant m$, and hence $M(\mathbf{s}^0)$ and $\pi_j(\mathbf{s}^0)$.

C. Obtain improved estimates s_j^1 by setting

$$s_j^1 = \pi_j(\mathbf{s}^0) \quad \text{for } 1 \leqslant j \leqslant m.$$

D. Return to step B with $\mathbf{s}^1$ replacing $\mathbf{s}^0$, etc.

E. Stop when the required accuracy has been achieved.

Note: Another way to write $\pi_j(\mathbf{s})$, which is useful if relatively few of the $\{A_i\}$ and $\{B_i\}$ are distinct, is

$$\pi_j(\mathbf{s}) = \left[\sum_A \xi_A I_A(j)\left(s_j \Big/ \sum_{k\in A} s_k\right) + \sum_B \eta_B\{1-I_B(j)\}\left(s_j \Big/ \sum_{k\in B} s_k\right)\right] \Big/ M(\mathbf{s}), \tag{3.5}$$

where ξ_A is the number of observations censored into the set A, η_B is the number truncated by the set B and $I_A(j)$ equals 1 if $[q_j, p_j] \subseteq A$ and is zero otherwise. Also, since

$$\sum_A \xi_A = \sum_B \eta_B = N,$$

$M(\mathbf{s})$ can be more simply expressed as

$$\sum_B \left\{\eta_B\left(\sum_{k\in B} s_k\right)^{-1}\right\}.$$

We now examine the equivalence of self-consistent and maximum likelihood estimation. From (2.2), we see that the log-likelihood is given by

$$L(\mathbf{s}) = \sum_{i=1}^{N}\left\{\log\left(\sum_{j=1}^{m}\alpha_{ij}s_j\right) - \log\left(\sum_{j=1}^{m}\beta_{ij}s_j\right)\right\}. \tag{3.6}$$

Consider the effect of increasing a particular component, s_j say, by a small positive amount ε and then dividing all the $\{s_k\}$, including $s_j+\varepsilon$, by $1+\varepsilon$ in order to keep the sum equal to unity. We let $d_j(\mathbf{s})$ denote the value of the derivative of L with respect to ε at $\varepsilon = 0$. Therefore

$$d_j(\mathbf{s}) = \frac{\partial}{\partial \varepsilon} L\left(\frac{s_1}{1+\varepsilon}, \ldots, \frac{s_j+\varepsilon}{1+\varepsilon}, \ldots, \frac{s_m}{1+\varepsilon}\right) \quad \text{at } \varepsilon = 0$$

$$= \frac{\partial L}{\partial s_j} - \sum_{k=1}^{m} s_k \frac{\partial L}{\partial s_k} \tag{3.7}$$

$$= \sum_{i=1}^{N}\left\{\left(\alpha_{ij} \Big/ \sum_{k=1}^{m}\alpha_{ik}s_k\right) - \left(\beta_{ij} \Big/ \sum_{k=1}^{m}\beta_{ik}s_k\right)\right\} \tag{3.8}$$

for $1 \leqslant j \leqslant m$. From (3.3) and (3.8) we have

$$\pi_j(\mathbf{s}) = \frac{s_j}{M(\mathbf{s})}\left\{d_j(\mathbf{s}) + \sum_{i=1}^{N}\left(\sum_{k=1}^{m}\beta_{ik}s_k\right)^{-1}\right\}. \tag{3.9}$$

However,

$$M(\mathbf{s}) = \sum_{i=1}^{N}\sum_{j=1}^{m}\left[\left(\alpha_{ij}s_j \Big/ \sum_{k=1}^{m}\alpha_{ik}s_k\right) + \left\{(1-\beta_{ij})s_j \Big/ \sum_{k=1}^{m}\beta_{ik}s_k\right\}\right]$$

$$= \sum_{i=1}^{N}\left(\sum_{k=1}^{m}\beta_{ik}s_k\right)^{-1}.$$

Substituting in (3.9), we obtain

$$\pi_j(\mathbf{s}) = \left\{1 + \frac{d_j(\mathbf{s})}{M(\mathbf{s})}\right\}s_j \quad (1 \leqslant j \leqslant m). \tag{3.10}$$

Now a necessary and sufficient condition for $\mathbf{s}$ to be an MLE is

$$\text{for each } j \quad \textit{either} \quad d_j(\mathbf{s}) = 0 \quad \textit{or} \quad d_j(\mathbf{s}) \leqslant 0 \quad \text{with } s_j = 0. \tag{3.11}$$

Thus from (3.10) and (3.11), we see immediately that the MLE $\hat{\mathbf{s}}$ satisfies $\pi_j(\hat{\mathbf{s}}) = \hat{s}_j$ for all j, and hence is self-consistent. Conversely, if the algorithm converges with limiting value $\tilde{\mathbf{s}}$, then $\tilde{\mathbf{s}}$ must satisfy (3.11). (A continuity argument shows we cannot have $d_j(\tilde{\mathbf{s}}) > 0$ with $\tilde{s}_j = 0$.)

Concerning convergence of the algorithm, we let $\mathbf{s}$ and $\mathbf{s}'$ be successive approximations where, by (3.10), $s_j' = [1+\{d_j(\mathbf{s})/M(\mathbf{s})\}]s_j$ for $1 \leqslant j \leqslant m$. Now by a Taylor series expansion we have

$$L(\mathbf{s}')-L(\mathbf{s}) = \sum_{j=1}^{m}(s_j'-s_j)\frac{\partial L}{\partial s_j}+O(\|\mathbf{s}'-\mathbf{s}\|^2)$$

$$\simeq \frac{1}{M(\mathbf{s})}\sum_{j=1}^{m}s_j d_j(\mathbf{s})\frac{\partial L}{\partial s_j} = \frac{1}{M(\mathbf{s})}\left\{\sum_{j=1}^{m}s_j\left(\frac{\partial L}{\partial s_j}\right)^2-\left(\sum_{j=1}^{m}s_j\frac{\partial L}{\partial s_j}\right)^2\right\}$$

$$= \frac{1}{M(\mathbf{s})}\sum_{j=1}^{m}s_j d_j^2(\mathbf{s}) \geqslant 0, \tag{3.12}$$

where we have used (3.7) and have neglected terms of second and higher order. Thus $L(\mathbf{s}') \geqslant L(\mathbf{s})$ with equality only if, for each j, either $s_j = 0$ or $d_j(\mathbf{s}) = 0$. Thus the algorithm converges monotonely, at least for $\mathbf{s}^0$ close enough to $\hat{\mathbf{s}}$ so that the higher order terms can indeed be neglected. Also it is clear that we must choose all $s_j^0 > 0$ (otherwise $s_j^k = 0$ for all k).

A maximum likelihood estimate $\hat{F}$ of F is given by

$$\hat{F}(x) = \begin{cases} 0 & \text{if } x < q_1, \\ \hat{s}_1+\hat{s}_2+\ldots+\hat{s}_j & \text{if } p_j < x < q_{j+1} \quad (1 \leqslant j \leqslant m-1), \\ 1 & \text{if } x > p_m, \end{cases}$$

and is undefined for $x \in [q_j, p_j]$ for $1 \leqslant j \leqslant m$. Therefore, when plotted, $\hat{F}$ consists of a series of $m+1$ horizontal lines of increasing heights with gaps in between, where the way in which increases occur is arbitrary. The variances and covariances of the non-zero $\{s_j\}$ are given by the inverse of the matrix of second derivatives of L with respect to the elements of $(s_1, s_2, \ldots, s_{m-1})$ corresponding to the non-zero elements of $\hat{\mathbf{s}}$. Thus estimates of the variance of $\hat{F}(x)$ can be calculated for $x \notin C$, from which approximate standard errors can be obtained for the height of each horizontal line.

4. Discussion

The self-consistency algorithm is automatic, simple to implement and is intuitively appealing. This contrasts with the direct but cumbersome constrained Newton–Raphson (NR) methods used by Peto (1973). Asano (1965) also used NR methods for a truncated multinomial. Hocking and Oxspring (1971) have proposed a similar algorithm for multinomial data subject to censoring without truncation. Also, upon reparametrization, it is possible to extend the method of Blight (1970) to this case, since the multinomial is in the exponential family of distributions.

For the problem of comparing two or more samples, each of which are subject to arbitrary censoring and/or truncation, $\hat{F}$ and L can be used to construct the analogue of the logrank test (Peto and Peto, 1972). Another possible application is to the regression methods of Miller (1974).

At a late revision of this paper, a referee kindly drew the author's attention to a recent unpublished manuscript by Dempster *et al.* (1976). In a very elegant and comprehensive theory of maximum likelihood with incomplete data, they develop the "EM" algorithm and its properties. The method described in Section 3 can be viewed as an example of an EM algorithm.

Acknowledgements

The author is grateful to Richard Peto and Nathan Mantel for some useful conversations and to the referees for their valuable comments.

This research was supported in part by grants DAHC04-73-C-0008, US Army Research Office—Durham, and N00014-75-C-0586, Office of Naval Research.

REFERENCES

ASANO, C. (1965). On estimating multinomial probabilities by pooling incomplete samples. *Ann. Inst. Statist. Math.*, **17**, 1–13.

AYER, M., BRUNK, H. D., EWING, G. M., REID, W. T. and SILVERMAN, E. (1955). An empirical distribution function for sampling with incomplete information. *Ann. Math. Statist.*, **26**, 641–647.

BLIGHT, B. J. N. (1970). Estimation from a censored sample for the exponential family. *Biometrika*, **57**, 389–395.

DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1976). Maximum likelihood from incomplete data via the EM algorithm. Research Report, Harvard University.

EFRON, B. (1967). The two sample problem with censored data. In *Proc. 5th Berkeley Symp. on Math. Statist. Prob.*, pp. 831–853. Berkeley: University of California Press.

HARTLEY, H. O. and HOCKING, R. R. (1971). The analysis of incomplete data. *Biometrics*, **27**, 783–823.

HOCKING, R. R. and OXSPRING, H. H. (1971). Maximum likelihood estimation with incomplete multinomial data. *J. Amer. Statist. Ass.*, **66**, 65–70.

KAPLAN, E. L. and MEIER, P. (1958). Nonparametric estimation from incomplete observations. *J. Amer. Statist. Ass.*, **53**, 457–481.

MANTEL, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports*, **50**, 163–170.

MILLER, R. G. (1974). Least squares regression with censored data. Technical Report, Stanford University.

PETO, R. (1973). Experimental survival curves for interval-censored data. *Appl. Statist.*, **22**, 86–91.

PETO, R. and PETO, J. (1972). Asymptotically efficient rank invariant test procedures. *J. R. Statist. Soc.* A, **135**, 185–206.

SUNDBERG, R. (1974). Maximum likelihood theory for incomplete data from an exponential family. *Scand. J. Statist.*, **1**, 49–58.

TURNBULL, B. W. (1974). Nonparametric estimation of a survivorship function with doubly censored data. *J. Amer. Statist. Ass.*, **69**, 169–173.